Concept Detection and Using Concept in Ad-hoc of Microblog Search

Hao Wu and Hui Fang
University of Delaware, Newark DE 19716, USA
haow@udel.edu, hfang@ece.udel.edu

Abstract:

We report our system and experiments in TREC 2012 microblog Ad-hoc task. Our goal is to return most relevant tweets to satisfy user's information needs which are represented by short keyword queries. In additional to the last year's temporal approach, we used Wikipedia pages to detect concepts of each query. And based on the concepts we detected, we did query expansion and concepts weighting separately. Both methods showed improvements comparing to baseline.

1 Introduction

Same as the last year's Ad-hoc task, the Tweets2011 collection is used which contains about 16 millions tweets over the period from Jan. 24th, 2011 to Feb. 8th, 2011. Each query is presented by short keywords and post time. Systems are required to return most relevant documents before the post time. Different from last year's Ad-hoc task which required returned tweets to be ranked by time, this year's Ad-hoc task requires returned tweets to be ranked by relevant score as state-of-art strategy. As a result, the system performance in this year is less sensitive to the number of returned tweets for each query.

Microblog search is a special kind of text search. Different from traditional text search whose document length is in a wide range, a tweet contains at most 140 characters. As a result, document length and query term frequency may not work as traditional information retrieval methods such as BM25. On the other hand, some other features such as temporal information or query concepts may play important roles. In additional to the last year's temporal approach, this year we explored the methods to detect the concepts of microblog search queries and we used the concept information to do query expansion and query term weighting. The experiment results showed improvements comparing to the baseline.

The rest of the paper is organized as following: we first explain our methods at section 2. Then we describe pre-processing and experiment results at section 3. Discussion is shown in section 4.

2 Method

Overview of our methods is shown as figure 1. For each query, we push them into Wikipedia pages to detect concepts (If there is corresponding wiki page, part of the original query is considered as one concept). For example, the query "British Government Cuts" is divided into two concepts "British Government" and "cut". Then based on the concepts we detect, we do query expansion and concepts weighting separately to get improved results.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
NOV 2012		**00-00-2012 to 00-00-2012**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Concept Detection and Using Concept in Ad-hoc of Microblog Search	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Delaware,Department of Electrical and Computer Engineering,Newark,DE,19716**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Presented at the Twenty-First Text REtrieval Conference (TREC 2012) held in Gaithersburg, Maryland, November 6-9, 2012. The conference was co-sponsored by the National Institute of Standards and Technology (NIST) the Defense Advanced Research Projects Agency (DARPA) and the Advanced Research and Development Activity (ARDA). U.S. Government or Federal Rights License

14. ABSTRACT
We report our system and experiments in TREC 2012 microblog Ad-hoc task. Our goal is to return most relevant tweets to satisfy user?s information needs which are represented by short keyword queries. In additional to the last year?s temporal approach, we used Wikipedia pages to detect concepts of each query. And based on the concepts we detected, we did query expansion and concepts weighting separately. Both methods showed improvements comparing to baseline.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**3**	

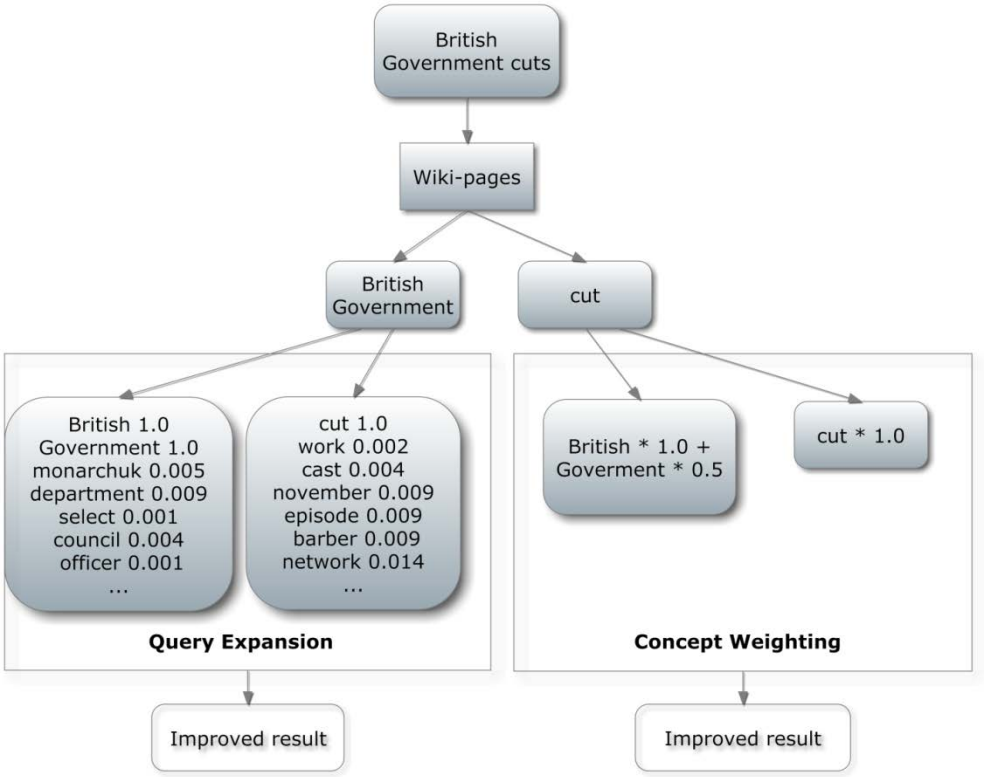

Figure 1 Overview of methods in microblog track 2012

2.1 Query Expansion

For each concept which has corresponding wiki page, we estimate the language model of the wiki page based on the frequency of each terms occurring in the page. Then we chose top 20 terms for each concept to add to the original query. The term weights in new query are determined by the probabilities of estimated language models.

2.2 Concepts Weighting

The basic idea of concept weighting is to favor tweets cover more concepts. Specifically, for query "British Government cuts" which divided into two concepts "British Government" and "cut", tweet contains "British cuts" or "Government cuts" should be ranked higher than tweet only talks about "British Government". To implement this strategy, we applied "discounted gain" for each concept. Specifically, the score gain of each concept is discounted as the number of terms increases:

$$S(Q,d) = \sum_{C_i \in Q} S(C_i,d) = \sum_{C_i \in Q} \sum_{t_{ij} \in C_i \ \& \ t_{ij} \in d} \frac{IDF(t_{ij})}{j}$$

For example, for query "British Government cuts". Tweet "British Government" will get full score for the term "British" but only half score for "Government" (IDF(British)*1.0+IDF(Government)*0.5), because it covers more than one terms in one concept. On the other hand, tweet "Government cut" will get full score for both term "Government" and "cut" (IDF(Government)*1.0+IDF(cut)*1.0), because both terms are the only one occurring in the tweet for each concept.

3 Experiment Results

3.1 Pre-processing and index building

We used "official twitter-corpus-tools corpus downloader" to download tweets with html output form. We downloaded the collection on June 8th, 2012 and got 10,887,718 tweets with 200 tags and 2992055 tweets with 301 tags.

We use character ASCII code to detect non-English tweets. In specifically, if a tweet contains character ASCII code larger than 127, it will be detected as non-English tweet and removed. In this way, we detected and removed 5,345,816 "non-English" tweets.

Link information in each tweet is not considered and removed. We left this part for future work.

We built the index with the Lemur toolkit. Porter stammer is applied and no stop words are removed. Tweets are treated as documents and they are sorted by their post time from old to new.

3.2 Experiment results

We submitted four runs:
UDInfoMBIDF is the baseline method which use query term IDF only
UDInfoMBTp applies the temporal method we used in 2011 TREC microblog track [1]
UDInfoMBEx applies concept based query expansion we describe in previous section.
UDInfoMBCW applies the concept weighting we describe in previous section

Result is shown in table below (highly relevant documents only):

	MAP	**R-prec**
UDInfoMBIDF (IDF only)	0.1040	0.1369
UDInfoMBTp(Temporal based ranking)	0.0960	0.1340
UDInfoMBEx (Concept based query expansion)	0.1161	0.1544
UDInfoMBCW(Concept Weighting)	0.1161	0.1562

4 Discussion

The temporal approach does not improve the performance. It may imply that our method of estimating the temporal information only by query term distribution is not enough. We plan to use more features to estimate the temporal information (e.g. distribution of retweets) and we leave it for future work.

On the other hand, Concept based query expansion and concept weighting improved the performance. It tells us that detecting concept may be useful in microblog search. We will do more work in this direction to get further improvement.

Reference

[1] H. Wu, H. Fang, "Time-Sensitive Weighting for Microblog Retrieval", in Proceeding of Text Retrieval Conference 2011.